DOCUMENT RESUME

ED 344 910 TM 018 233

AUTHOR Powell, Z. Emily
TITLE Test Anxiety and Test Performance Under Computerized Adaptive Testing Methods.
PUB DATE Apr 92
NOTE 34p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, April 20-24, 1992).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Adaptive Testing; Comparative Testing; *Computer Assisted Testing; *Graduate Students; Higher Education; *Mastery Tests; Scores; Student Attitudes; *Test Anxiety; Test Items; Test Results
IDENTIFIERS *Psychological Influences; *Testing Effects

ABSTRACT

Little research exists on the psychological impacts of computerized adaptive testing (CAT) and how it may affect test performance. Three CAT procedures were examined, in which items were selected to match students' achievement levels, from the item pool at random, or according to student choice of item difficulty levels. Twenty-four graduate students (5 males and 19 females) at Indiana University (Richmond) were randomly assigned to one of six testing orders formed by the three mastery test approaches and by blocking on native and non-native speakers. While at a computer, students received a description of adaptive testing methods and then a 20-item pre-test anxiety measure. Right after completing each test, students responded to a 10-item in-test anxiety scale, ranked their preferences among tests, and evaluated their performance on each of the tests. No statistically significant mean differences were found among mean student achievement scores or among in-test anxiety means under the three adaptive testing methods. Students reporting higher anxiety scored significantly higher in the matched-selection test. Those preferring the matched-selection and self-selection tests the most were less anxious during those tests. Instead of actual performance, students' perceptions of how well they did were significantly correlated with preference rankings for the tests. The matched-selection tests required significantly fewer items to reach decisions than did the random-selection tests. Ten tables, 1 figure, and a 19-item list of references are included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Zen-Hsiu Emily Powell

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED344910

Test Anxiety and Test Performance

Under Computerized Adaptive Testing Methods

Z. Emily Powell

Division of Education, Indiana University East

Paper for Presentation at the Annual Conference of the
American Educational Research Association
San Francisco

April, 1992

TM018233

BEST COPY AVAILABLE

ABSTRACT

Research on computerized adaptive testing (CAT) has shown its potential to estimate student achievement accurately and efficiently. However, there is a paucity of research on the psychological impacts of CAT and how it may, consequently, affect test performance. Three kinds of computerized adaptive testing procedures were examined, in which items were selected in one of three ways: (a) to match students' estimated achievement levels, (b) from the item pool at random, and (c) according to student choice of item difficulty levels.

Twenty-four graduate students were randomly assigned to one of six possible testing orders formed by the three adaptive tests while blocking on native and nonnative speakers. While at a computer, students first received a description of adaptive testing methods, followed by a 20-item pre-test anxiety measure. The three adaptive tests were then given in the assigned random order. Immediately after each adaptive test, students responded to a 10-item in-test anxiety scale. Finally, they ranked their preferences and evaluated their performance on each of the three tests.

No statistically significant differences were found among mean student achievement scores, nor among in-test anxiety means, under the three adaptive testing methods. Students who reported higher pre-test anxiety were found to score significantly higher in the matched-selection test. Students who preferred the matched-selection and self-selection tests the most tended to be significantly less anxious during those tests. In addition, instead of students' *actual* test performance, it was their *perception* of how well they did that was significantly correlated with their preference rankings for the three tests. Also, the matched-selection tests required significantly fewer items to reach decisions than did the random-selection tests.

OVERVIEW

Extant research on computerized adaptive testing has shown its potential for making accurate and efficient mastery/nonmastery classifications of testees (Frick, 1989; Frick et al., 1989; Plew, 1989). In an adaptive mastery testing setting, an item whose difficulty most closely matches a testee's achievement level is selected from the item bank and administered. Procedurally, a more difficult item is selected following a correct response, while an incorrect response is followed by an easier item. The test is terminated whenever a classification decision can be made accurately.

After years of investigation and research, Frick (1989) identified characteristics associated with two approaches to computerized adaptive testing that may have hindered their applications by classroom practitioners. The first approach is based on Wald's (1947) *Sequential Probability Ratio Test*. The SPRT administers test items at random, and updates the probability of a testee's being a master or a nonmaster according to the number of correct and incorrect responses. The updated ratio of mastery to nonmastery probabilities is compared with some predetermined accuracy values after each response, and the test is terminated as soon as a decision can be made. Even though the SPRT has been empirically shown to be efficient and accurate when used conservatively, Frick (1989) indicated that:

1. The SPRT does not consider item characteristics (e.g., difficulty, discrimination, guessing factors, etc.).

2. The SPRT does not select items to match the testee's achievement level.

Considerable research has been devoted to the second approach to computerized adaptive testing, termed *Item Response Theory* (Lord & Novick, 1968). Instead of making

discrete binary categories as does the SPRT, the IRT views the testee's achievement level as a continuum. After each item is answered, the current achievement level of the testee is estimated, and also, a confidence interval of the estimated achievement is computed. The obtained interval is checked against a predetermined mastery theta (θ) value. If the lower bound of the interval exceeds the mastery theta, the test is terminated and a mastery decision is reached; or if the upper bound of the interval is lower than the mastery theta, the test is terminated and a nonmastery decision is made. If neither of the cases above is true, the test continues and the item that has the difficulty level nearest to the estimated achievement level is chosen next (Kingsbury & Weiss, 1979; Frick, 1989; Plew, 1989).

In a historical review reported by Plew (1989) it was showed that the IRT can reduce the test length significantly, and it can also perform accurate achievement estimates. However, both Frick (1989) and Plew (1989) signified that:

1. The IRT Requires a large number of prior administrations, around 200 to 1000 depending on the model used, in order to acquire accurate item parameter calibrations.

2. The IRT involves complex statistical computations which are difficult to be comprehended by non-measurement specialists.

To address these disadvantages related to computerized adaptive testing strategies, Frick et al. (1989) and Plew (1989) corporately modified the SPRT so that item characteristics are considered, and the computations of achievement estimation are much simpler than the IRT. The new approach is named the *Expert Sequential Probability Ratio Test*, abbreviated as *EXSPRT*. According to the EXSPRT method, item difficulty

and discrimination parameters are calibrated based on past administrations--50 administrations appear to be sufficient (Frick et al., 1989; Plew, 1989). During the test administration the EXSPRT method continuously uses Bayes' Theorem to estimate the achievement level of the testee, which is defined as the respective probabilities of being a master and a nonmaster. As soon as the probability of either alternative (mastery/nonmastery) is found to be significantly greater than the other, a corresponding decision is made and the test is terminated (Frick et al., 1989; Plew, 1989). The EXSPRT method can be incorporated with different item selection methods. Three procedures were investigated in the present study. The first procedure uses an intelligent algorithm to select items matching the testee's estimated achievement level, termed *EXSPRT-I.*

One of the criticisms facing the algorithm-guided adaptive method is its predictability on item selection. The nature of adaptive testing algorithms is to take the number of correct/incorrect responses that have been given thus far to estimate the testee's current achievement level and then select the most closely matched item. Items selected in the algorithm-guided test are based on the statistical property of each item, instead of on the overall content presentation. Due to the unchanged statistical properties, the sets of items selected will be *identical* if a testee responds to each item consistently in different algorithm-guided tests. In order to sample test items across the content area, the present study includes the randomized item selection method as one of the treatments to be studied. The method is termed *EXSPRT-R*, which performed relatively efficiently and accurately in the re-enactment study by Frick et al. (1989).

An issue in educational computing research has been whether testing guided by testees themselves will result in the same or better performance and, at the same time, reduce test anxiety when compared to either the program-controlled or the conventional test. Rocklin and O'Donnell (1979) showed that, while given the control over item selection, testees were able to make appropriate decisions on the difficulty level of test items. To further verify this notion, the present study proposed a self-adaptive test (*EXSPRT-S*) that allows testees, rather than the adaptive algorithms, to choose item difficulty categories from which to receive test items.

SIGNIFICANCE OF THE STUDY

There were two main interests in the present study. The first is to investigate the psychological impact of the three computerized adaptive mastery tests. Much of the research on adaptive testing has been focused on its psychometric characteristics (Betz & Weiss, 1973, 1974; Vale & Weiss, 1975; Frick, 1989). There has been little research on testees' affective reactions during an adaptive test. The findings of such research might lead to improvement in the psychological environment of adaptive tests.

Some studies (Hambleton & Traub, 1974; Rocklin, 1985; Rocklin & O'Donnell, 1987) have shown that item difficulty affects test performance. The second interest of the present study is to examine test performance under the three computerized adaptive mastery tests. Test performance on the self-guided adaptive test would be compared with algorithm-guided and random selection adaptive tests. The results should help us understand further how testees perform during a self-adaptive testing condition. In

addition, the results would reveal whether an algorithm-guided testing method can significantly improve test performance when compared with a non-guided situation, i.e., random item selection.

METHOD

The present experimental study, served as a diagnostic test of student learning about functions of computers, and was incorporated in the existing curriculum of a basic computer course. The whole study was carried out completely in a real life classroom setting on AT&T personal computers. Color monitors were used for display, and floppy disks for test administration and data storage. Prior to the test, the instructor of the course introduced the method of computerized adaptive mastery testing in the class to ensure that testees were aware of the difference between the adaptive tests and the conventional test.

Subjects

Subjects of the present study were 24 graduate students enrolled in an introductory course on using computers in education at Indiana University. The majority of the class were first-year students on campus, with half being native speakers and the other half nonnative speakers, i.e., students from other countries. Five of the 24 students were male. On a self-report record, students reported a wide range of computer experience, varying from one month to five years. Note: Since the present study was conducted in a testing setting, subjects are referred to as "testees" later in this report.

Design

The present study used randomized block design with repeated measures (Kirk, 1982). In view of a possible language disadvantage of new international students, testees were divided into two groups, native and nonnative, and then from these two groups testees were assigned randomly in equal number to six groups. Thus, each of the six groups had the same number of native and international students, i.e., two. The purpose of forming the six groups for the three levels of treatment was to control a possible order effect. These three levels of treatment were the three different computerized adaptive mastery testing methods, viz., EXSPRT-I, EXSPRT-R, and EXSPRT-S. Procedurally, each testee took all three tests in one of six different orders, e.g., EXSPRT-I, EXSPRT-R, EXSPRT-S; EXSPRT-I, EXSPRT-S, EXSPRT-R; EXSPRT-R, EXSPRT-I, EXSPRT-S; etc.

Independent Variables

As mentioned above, three computerized adaptive mastery testing procedures were investigated in this study: (a) EXSPRT-I, (b) EXSPRT-R, and (c) EXSPRT-S. Each of these three adaptive mastery tests employed a different procedure to select items from the item bank; whereas they all utilized the EXSPRT discrete Bayesian estimation method to update the testee's current performance after each response; and terminated the test as soon as a decision about the testee's mastery/nonmastery status could be accurately made. Since the method of item selection was the sole difference among the three adaptive mastery testing procedures, only this prospect is discussed below.

EXSPRT-I. The "intelligent" algorithm embedded in the EXSPRT-I test enabled this adaptive method to select the best item to match the testee's achievement level (Frick et al., 1989; Plew, 1989). The algorithm consistently calibrated item characteristics (i.e., item difficulty, item discrimination) for each item remaining in the bank after each testee response. The EXSPRT intelligent item selection method always attempted to select the item that was highly discriminative and that had a difficulty level near the testee's estimated achievement level. An item's difficulty level is conceptualized as the probability of an item's being answered incorrectly by the whole sample, including both masters and nonmasters. Thus, the higher difficulty index an item has, the more difficult it is to the testees. Item discrimination is defined as the difference between the probability of an item's being answered correctly by a master and that by a nonmaster. A more discriminative item is considered to be a better item.

EXSPRT-R. Instead of selecting items based on the testee's performance, the EXSPRT-R test selected items from the bank at random. Since the statistical property of an item did not affect its chance of administration; the EXSPRT-R test should result in better overall content representation than the EXSPRT-I.

EXSPRT-S. In the EXSPRT-S test, the testee was asked to select a difficulty category, and from the category the most discriminative item was then administered. Items in the bank were evenly divided into five difficulty categories: (a) very easy, (b) easy, (c) medium, (d) difficult, and (e) very difficult. Within each difficulty category, items were administered in order from the most to the least discriminating. For instance, if the "medium" category was chosen the first time, the most discriminative item in this

category was administered; and if this category was chosen the second time, the next most discriminative item was administered and so forth.

Again, the length of each of the three adaptive mastery tests depended on how well (or how poorly) the testee performed during the test. A test might be very short if the testee consistently scored correctly on the test, or in the other way, incorrectly; whereas a test might last longer if a series of mixed correct and incorrect responses were given.

Dependent Variables

As it was mentioned earlier, the testee's test anxiety and performance were the two main interests of the present study. In addition, because the adaptive methods were new to the testees, other aspects of their affective reactions such as preferences for the different testing procedures and how well they thought they did on the tests were also solicited.

Test anxiety. Two kinds of test anxiety were measured in this study: a) pretest anxiety, and b) in-test anxiety. The STAI A-State scale (Spielberger et al., 1970) was given to testees prior to the adaptive tests to measure their pretest anxiety. Testees were asked to respond to each of 20 statements by rating themselves on a four-point scale: "Not At All", Somewhat", "Moderately So", and "Very Much So". The range of possible scores for the A-State scale varies from a minimum score of 20 to a maximum score of 80.

In-test anxiety was measured using the modified Test Anxiety Inventory (Speilberger et al., 1980). The original TAI scale consists of 20 items designed to

measure test anxiety before, during, and after the test. Since this study was primarily concerned about test anxiety during the tests, 10 of the original 20 items that specifically ask testees how they feel during a test were chosen and used. Immediately following each adaptive test, testees were asked to respond to each of 10 statements by rating themselves on a four-point scale: "Almost Never", "Sometimes", "Often", and "Almost Always". The range of possible scores for the modified TAI scale varies from a minimum of 10 to a maximum of 40.

Test performance. In this study, testees' test performance was estimated using the EXSPRT discrete Bayesian method. At the beginning of the test, the EXSPRT discrete Bayesian method assumed each testee was an average performer, i.e., with a 50% of probability of mastering the learning objectives in the test. During the test, the probabilities of the testee's being a master or nonmaster were consistently re-estimated after each response. When either probability became significantly greater than the other, the test was terminated and a decision about the testee's mastery was then made accordingly. The testee was judged either a master or nonmaster. Note: Prior to the tests, the mastery cutoff was set to be 85%, which conventionally means that out of the total 85 items in the bank anyone who has correctly answered 72 items or more is classified as a master and the others as nonmasters. However, in the adaptive mastery test not all of the items in the bank were administered. Also, unlike the conventional test, each item in the bank was weighed differently according to its associated characteristics such as item difficulty and item discrimination.

Preference rankings. Immediately after the three adaptive tests, testees were asked to rank the tests in order of their preference. The three color-coded adaptive tests were listed on the computer screen from left to right according to their administration order. Testees were first asked to choose the one they liked most, then the one they liked next. If a test was chosen more than once, a message was displayed to inform testees that the test had already been picked, and that they should make their choice from the remaining tests.

Self-evaluation. Before the test results were given, testees were asked to evaluate their test performance on the three adaptive tests. The three color-coded adaptive tests were listed in parallel on the computer screen in their administration order. Testees were first asked to choose the test on which they thought they did best, then the test on which they did next best. If a test was chosen more than once, a message was displayed to tell testees to make they choice from the remaining tests.

Procedures

At the beginning of the tests, each testee was seated in front of a computer and instructed how to start the tests stored on a floppy disk provided with the computer. The testee first read eight screens of descriptions about how to operate the program, objectives of the tests, consequences of test performance, and the purpose of the embedded test anxiety scales. Also, the basic concept of the adaptive mastery testing method was restated.

Next, testees were asked to respond to each statement of the 20-item STAI A-State Scale (Spielberger et al., 1970) according to how they felt right at that moment by

choosing 1 to 4 representing from "Not At All" to "Very Much", respectively. Each statement was displayed on the screen separately, one after another.

Following the STAI A-State Scale, testees received the three adaptive mastery tests in the order assigned to them. Two screens containing: (a) a description of the computerized adaptive mastery testing method to follow, and (b) an explanation of how long the test would last were given before each adaptive test. Not to confuse testees with uncommon terms, the EXSPRT-I method was referred to as "matched selection", the EXSPRT-R as "random selection" and the EXSPRT-S as "self selection".

In addition, to minimize possible performance discrepancy due to misunderstanding of the adaptive method, testees were required to correctly answer five questions regarding the current method before they were allowed to proceed with the test. However, additional oral explanations were given to any individual who failed any of the five questions more than three times. During the present study, there was only one student who needed such help.

Immediately after each adaptive test, the modified 10-item Test Anxiety Inventory (Spielberger et al., 1980) was given. Testees were instructed to respond to each item based on their feelings specifically during the adaptive test they had just taken.

Finally, after having finished all three adaptive tests and the test anxiety scale following each test, testees were asked to rank the three adaptive mastery tests in order of their preference, and also to rank how well they thought they did on each test. Not until then were testees informed how well they actually did on the tests.

RESULTS

One of the two primary questions posed by the present study was whether there was statistically significant difference in testees' test performance under the three computerized adaptive mastery testing procedures. In other words, the present study was concerned whether the item selection procedures would cause discrepancy in test performance. As depicted in Table 1, 11 of the 24 testees were classified as masters in the EXSPRT-I test, whereas 13 and 14 in the EXSPRT-R and EXSPRT-S tests, respectively. The chi-square test of association was computed to determine whether test performance differed among the three tests. The yielded value of chi-square resulted in a significance level of $p < .7$, which was greater than .05. Therefore, it was concluded that there was no statistically significant difference in test performance among the three tests.

Insert Table 1 about here.

The other main interest of the present study was to determine whether the method of item selection affected the level of in-test anxiety. The results, which were computed using an analysis of variance with repeated measures and detailed in Table 2, showed that the F test for in-test anxiety under the three adaptive mastery testing methods was not significant at the .05 level ($F = .93$, $df = 2,44$). That is to say that the testee experienced a similar level of in-test anxiety during all three tests.

Insert Table 2 about here.

Additional data analyses were carried out to investigate the relationships between different aspects of the testee's affective reactions to the adaptive testing procedures. First, the testee's pre-test anxiety was correlated with test performance to reveal the relationship. Table 3 showed that the Pearson correlation coefficient of pre-test anxiety and test performance under the EXSPRT-I test was significant. This suggested that those testees who felt more anxious before the tests did better in the EXSPRT-I test. However, there was no correlation between pre-test anxiety and test performance under the EXSPRT-R or EXSPRT-S test.

Insert Table 3 about here.

Pearson correlation coefficients of in-test anxiety and test performance under the three tests were also computed to determine whether there was a significant relationship between the two. The results, shown in Table 4, indicated that testees' in-test anxiety level was not related to how well they performed on the tests.

Insert Table 4 about here.

To investigate whether other factors might have affected test performance, testees'

preference rankings for the three tests were correlated with test performance. A nonparametric correlation statistic, Kendall's tau-b, was computed to reveal their relationship. The obtained Kendall correlation coefficients and the associated significance levels are reported in Table 5. None of the correlation coefficients reached significance at the .05 level. This suggested that how the testees felt about the tests did not affect their test performance.

Insert Table 5 about here.

Further, Table 6 lists Kendall's tau-b, the nonparametric correlation coefficients of testees' preference rankings and their in-test anxiety during the three tests. The coefficients for the EXSPRT-I and EXSPRT-S tests were significant at .05 level. This indicated that testees who were more anxious during these two tests also preferred the tests least. No such relationship was found for the EXSPRT-R test.

Insert Table 6 about here.

Since the method of adaptive mastery testing procedures was new to testees, an interesting question regarding whether they knew how well they did on the tests was raised. The nonparametric correlation coefficients of testees' self evaluations of test performance under the three tests and their actual test performance were listed in Table 7. The results showed that there was no significant relationship between the two for any

of the three tests. This indicated that in fact testees did not know how they actually performed on the tests.

Insert Table 7 about here.

Thus, further analyses were computed to reveal whether the testee's self-evaluation was associated with in-test anxiety, and the preference ranking of the tests. The results are listed in Tables 8 and 9. It appears that how testees perceived their test performance was not related to their in-test anxiety levels, however, it was significantly associated with their preference rankings of the tests. This finding and the previous finding regarding test performance and the preference ranking suggested that testees' *perceptions* of their test performance rather than their *actual* test performance might have affected their preference rankings for the tests.

Insert Tables 8 and 9 about here.

Numerous studies have proven that the adaptive testing method is more efficient than the conventional test. The results of Frick's (1990) simulation study showed that the EXSPRT-I testing method required significantly fewer items than other adaptive testing methods, including the EXSPRT-R method. To verify this finding, the ANOVA of the total numbers of items used in each test was computed and detailed in Table 10. The results indicated that there was a significant difference among test lengths under the

three adaptive mastery testing methods. Tukey's pairwise test was computed to reveal the significantly different pair(s). It was found that the mean test length required for the EXSPRT-I test was significantly shorter than that for the EXSPRT-R test.

Insert Table 10 about here.

SUMMARY AND DISCUSSION

The main purpose of the present study was to investigate whether different methods of item selection in computerized adaptive mastery tests affected test performance, and in-test anxiety. As depicted in Figure 1, three hypothesized relationships were: (a) Relationship A--different methods of item selection resulted in different levels of test performance, (b) Relationship B--different methods of item selection altered the levels of in-test anxiety, and (c) Relationship C--different levels of in-test anxiety resulting form different methods of item selection consequently affected test performance.

Insert Figure 1 about here.

The results did not support the hypothesized Relationship A because estimates of testees' mean achievement under the three adaptive mastery testing methods were not significantly different. However, one method may be more efficient than another by

virtue of "better" items being selected. As in the present study, the EXSPRT-R test required 1/3 of the items in the item bank whereas the EXSPRT-I test required only 1/6 of the items (with the EXSPRT-S test in between).

Also, contrary to predictions, the results of in-test anxiety did not support the proposed Relationship B since no statistically significant difference was found in testees' in-test anxiety under the three adaptive mastery tests. Two suppositions are suggested for this finding. First, the test results did not have negative consequences on the final course grade. Because the three adaptive tests given in the present study were used for the diagnostic purpose, testees who did not master all three adaptive tests were given a second chance to make up the difference on a parallel posttest. Previous studies have shown that the perceived importance of test results is highly related to test performance (Raynor, 1970; Simons & Bibb, 1974). Thus, it is presumed that if no second chance had been available, and as a result of fearing possible negative consequences of the test results from the adaptive tests, testees might have reflected different levels of in-test anxiety.

Second, short intervals between repeated measures of in-test anxiety might have hindered possible significant results. The interval between the in-test anxiety measures was about 10 to 15 minutes depending on how individuals performed on the previous adaptive test and how they responded to five preview questions regarding the testing method. Because measurements of in-test anxiety occurred very close to one another, a carry-over effect might have resulted. In other words, anxiety derived from one test might be carried over to the next because of the short interval between the tests.

Because Relationship C is an extension of Relationship B, which was found to be untenable, there is no grounding to support Relationship C. Further data analysis showed that there was no relationship between testees' self-evaluations of test performance and their actual test performance under the three tests. This finding suggested that testees did not really know how well they did on the tests. However, this phenomenon is understandable. As stated earlier, one of the main differences between conventional tests and adaptive tests is their method of achievement estimation. Unlike the conventional test, the simple number of correct responses is no longer the sole factor used to reach conclusions in an adaptive test. Item characteristics are other important factors for the estimation of achievement level. It was assumed that this was the reason why testees were unable to estimate accurately how well they had performed on the three adaptive tests.

Additionally, the results showed that testees who experienced lower levels of in-test anxiety also liked the tests better under the EXSPRT-I and EXSPRT-S testing procedures. Though no relationship was found between testees' *actual* test performance and their preference rankings for the tests, their *perceptions* of how well they did on the tests did correlate significantly with the preference rankings of the three tests. It suggests that testees who thought they did best on a given test liked that test best. Also, testees' self-evaluations for their performance on the EXSPRT-I and EXSPRT-S tests were found to be somewhat correlated with their in-test anxiety. Testees who thought they did best on the EXSPRT-I and EXSPRT-S tests experienced the least anxiety during that test.

Based on the findings of the present study and the investigator's suppositions, two things are recommended for further research studies which encompass the same interests. First, recruit more samples so subjects are measured in only one testing condition to avoid any possible carry-over effects. Second, ensure that testees will treat the tests as "real". Since the importance of test results is highly related to test performance (Raynor, 1970; Simons & Bibb, 1974), it is critical to assure that testees try to perform as well as they can to reveal their true achievement levels.

REFERENCES

Betz, N. & Weiss, D. (1973). *An empirical study of computer administered two-stage ability-testing* (Research Report 73-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Betz, N. & Weiss, D. (1974). *Simulation studies of two-stage ability testing* (Research Report 74-4). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Frick, T. (1989). Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research*, *5*(1), 89-114.

Frick, T. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research*, *6*(4), 479-513.

Frick, T., Plew, G. & Luk, H. (1989, March). *EXSPRT: An expert systems approach to computer-based adaptive testing.* Paper presented at the Annual Conference of the American Educational Research Association. San Francisco.

Hambleton, R. & Traub, R. (1974). The effects of item order on test performance and stress. *Journal of Experimental Education*, *43*(1), 40-46.

Hopkins, D. & Glass, G. (1978). *Basic statistics for the behavioral sciences.* Englewood Cliffs, NJ: Prentice-Hall.

Kingsbury, G. & Weiss, D. (1979). *An adaptive strategy for mastery decisions* (Research Report 79-5). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Methods Program.

Kirk, R. (1982). *Experimental design: Procedures for the behavioral sciences.* Belmont, CA: Wadsworth.

Lord, F. & Novick, M. (1968). *Statistical theories of mental test scores.* Reading: Addison-Wesley.

Plew, G. (1989). *An empirical investigation of major adaptive testing methodologies and an expert systems approach.* Unpublished doctoral dissertation, Indiana University, Bloomington.

Raynor, J. (1970). Relationships between achievement-related motives, future orientation, and academic performance. *Journal of Personality and Social Psychology*, *15*(1), 28-33.

Rocklin, T. (1985). Interactive effects of test anxiety, test difficulty, and feedback: Implications for ability testing. In H. van der Ploeg, R. Schwarzer, & C. Spielberger (Eds.), *Advances in test anxiety research* (Vol. *4*, pp. 79-88). Hillsdale, NJ: Erlbaum.

Rocklin, T. & O'Donnell, A. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology*, *79*(3), 315-319.

Simons, R. & Bibb, J. (1974). Achievement motivation, test anxiety, and underachievement in the elementary school. *Journal of Educational Research*, *67*(8), 366-369.

Spielberger, C., Gonzalez, J., Taylor, C. & Anton, W. (1980). *Test attitude inventory.* Palo Alto, CA: Consulting Psychologists Press.

Spielberger, C., Gorsuch, R. & Lushene, R. (1970). *The state-trait anxiety inventory.* Palo Alto, CA: Consulting Psychologists Press.

Vale, D. & Weiss, D. (1975). *A simulation study of stradaptive ability testing* (Research Report 75-6). Minneapolis: University of Minnesota, Department of Psychology, Psychometric Program.

Wald, A. (1947). *Sequential analysis.* New York: Wiley.

Table 1
Frequency of the EXSPRT Discrete Decisions Under Each Computerized Adaptive Mastery Testing Method

Decision	Adaptive Testing Method		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
Mastery	11	13	14
Nonmastery	13	11	10
Chi-square = .78	n.s.		
$r_{\phi IR}$ = .34	n.s.		
$r_{\phi IS}$ = .61	***		
$r_{\phi RS}$ = .58	***		

Note: Due to the repeated measures, the observations in this contingency table were not statistically independent. Consequently, the probability of a type-I error of this chi-square test was seriously affected (Hopkins & Glass, 1978).

***$p < .005$.

Table 2

ANOVA Results of In-test Anxiety Under Each Computerized Adaptive Mastery Testing Method

Source	*SS*	*df*	*MS*	*F*	Sig. of *F*	
Between blocks	1905.11	23				
A(Language)	133.39	1	133.39	1.66	.211	n.s.
Blocks w. A	1771.72	22	80.53			
Within blocks	178.00	48				
B(Adaptive methods)	7.19	2	3.60	.93	.402	n.s.
AB	.53	2	.26	.07	.934	n.s.
B x blocks w. A	170.28	44	3.87			
Total	2083.11	71				

Mauchley Sphericity Test, $W = .800$, $p = .096$.

Table 3

Pearson Correlations of Pre-test Anxiety and Test Performance

	Test Performance		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
	r (*p*)	*r* (*p*)	*r* (*p*)
Pre-test Anxiety	.549*** (.005)	-.038 (.860)	.192 (.368)

***$p < .005$.

Table 4

Pearson Correlations of In-test Anxiety and Test Performance

	Test Performance		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
In-test Anxiety	*r* (*p*)	*r* (*p*)	*r* (*p*)
EXSPRT-I	.299 (.156)	--	--
EXSPRT-R	--	-.139 (.516)	--
EXSPRT-S	--	--	-.057 (.792)

Table 5

Nonparametric Kendall Correlations of the Preference Ranking and Test Performance

	Test Performance		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
Preference Ranking	*tau-b* (*p*)	*tau-b* (*p*)	*tau-b* (*p*)
EXSPRT-I	.057 (.773)	--	--
EXSPRT-R	--	.166 (.400)	--
EXSPRT-S	--	--	-.063 (.752)

Note. The preference ranking was ranged from 1 to 3, 1 = "Like Most", 2 = "Like Next", and 3 = "Like Least". Greater Bayesian estimates indicated higher test scores.

Table 6

Nonparametric Kendall Correlations of the Preference Ranking and In-test Anxiety

	In-test Anxiety		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
Preference Ranking	*tau-b* (*p*)	*tau-b* (*p*)	*tau-b* (*p*)
EXSPRT-I	.383* (.035)	--	--
EXSPRT-R	--	-.095 (.596)	--
EXSPRT-S	--	--	.379* (.035)

Note. The preference ranking was ranged from 1 to 3, 1 = "Like Most", 2 = "Like Next", and 3 = "Like Least". Higher in-test anxiety scores indicated greater anxiety levels.

*$p < .05$.

Table 7

Nonparametric Kendall Correlations of Self-evaluation and Test Performance

	Test Performance		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
Self Evaluation	*tau-b* (*p*)	*tau-b* (*p*)	*tau-b* (*p*)
EXSPRT-I	.069 (.728)	--	--
EXSPRT-R	--	.193 (.326)	--
EXSPRT-S	--	--	-.032 (.874)

Note. Self-evaluation was ranged from 1 to 3, 1 = "Did Best", 2 = "Next Best", and 3 = "Did Worst". Greater Bayesian estimates indicated higher test scores.

Table 8

Nonparametric Kendall Correlations of Self-evaluation and In-test Anxiety

	In-test Anxiety		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
Self Evaluation	*tau-b* (*p*)	*tau-b* (*p*)	*tau-b* (*p*)
EXSPRT-I	.351 (.053)	--	--
EXSPRT-R	--	.152 (.394)	--
EXSPRT-S	--	--	.318 (.078)

Note. Self-evaluation was ranged from 1 to 3, 1 = "Did Best", 2 = "Next Best", and 3 = "Did Worst". Higher in-test anxiety scores indicated greater anxiety levels.

Table 9

Nonparametric Kendall Correlations of Self-evaluation and the Preference Ranking

	Preference Ranking		
	EXSPRT-I	EXSPRT-R	EXSPRT-S
Self Evaluation	*tau-b* (*p*)	*tau-b* (*p*)	*tau-b* (*p*)
EXSPRT-I	.739*** (.000)	--	--
EXSPRT-R	--	.584*** (.002)	--
EXSPRT-S	--	--	.481* (.011)

Note. Self-evaluation was ranged from 1 to 3, 1 = "Did Best", 2 = "Next Best", and 3 = "Did Worst". The preference ranking was also ranged from 1 to 3, 1 = "Like Most", 2 = "Like Next", and 3 = "Like Least".

*$p < .05$. ***$p < .005$.

Table 10

ANOVA Results of Test Length Under Each Adaptive Testing Method

Source	*SS*	*df*	*MS*	*F*	Sig. of *F*	
Between blocks	8110.21	23				
A(Language)	141.68	1	141.68	.39	.538	n.s.
Blocks w. A	7968.53	22	362.21			
Within blocks	8742.66	48				
B(Adaptive methods)	2847.58	2	1423.79	10.67	.000	***
AB	24.19	2	12.10	.09	.913	n.s.
B x blocks w. A	5870.89	44	133.43			
Total	16852.87	71				

Mauchley Sphericity Test, $W = .845$, $p = .092$.

Tukey statistics: ${}_{.95}q_{3,44} \approx 3.44$

$q_{1R} = 3.90^{*}$
$q_{RS} = 2.78$
$q_{1S} = 1.12$

$^{*}p < .05$. $^{***}p < .005$.

Figure 1

Relationships of Methods of Item Selection, Test Performance and In-test Anxiety